IN THE UNITED STATES DISTRICT COURT
FOR THE MIDDLE DISTRICT OF TENNESSEE
NASHVILLE DIVISION

CONCORD MUSIC GROUP, INC., ET AL., *Plaintiffs*, v. ANTHROPIC PBC, *Defendant*.	Case No. 3:23-cv-01092 Chief Judge Waverly D. Crenshaw, Jr. Magistrate Judge Alistair Newbern

DECLARATION OF STEVEN R. PETERSON, PHD
IN SUPPORT OF DEFENDANT'S OPPOSITION TO
PLAINTIFFS' MOTION FOR PRELIMINARY INJUNCTION

TABLE OF CONTENTS

Expert Report of Steven R. Peterson

I. QUALIFICATIONS AND ASSIGNMENT

A. Qualifications

1. My name is Steven R. Peterson. I am an Executive Vice President at Compass Lexecon. Compass Lexecon is an economics consulting firm that specializes in the economics of competition, finance, and regulation, among other areas. I received my A.B. in economics from the University of California, Davis, in 1987 and my Ph.D. in economics from Harvard University in 1992. While at Harvard, my areas of specialization were economic theory and industrial organization. Industrial organization is the study of the interactions of larger firms that can strategically influence the markets in which they operate. I have also served as an adjunct faculty member in the Department of Economics at Northeastern University where I taught courses on the economics of antitrust, regulation, and public policy.

2. I have consulted on the economics of antitrust and competition, mergers, estimation of damages, valuation, regulation, and public policy. I have extensive experience with the policy and economics of copyrights. I testified on behalf of the Radio Music License Committee ("RMLC") on issues of market definition and market power in *Radio Music License Committee v. SESAC, Inc., SESAC, LLC, and SESAC Holdings, Inc.*, and I testified on public performance license rates for radio in the subsequent arbitration proceedings. I addressed similar issues in RMLC's litigations against Global Music Rights ("GMR") and Broadcast Music, Inc. I also testified before the United States Copyright Royalty Judges, Library of Congress, regarding appropriate rates for digital public performances of sound recordings in the *Web IV* (on behalf of the National Association of Broadcasters) and *Web V* (on behalf of Google) proceedings. I submitted comments on behalf of the National Association of Broadcasters addressing the Department of Justice's review of the American Society of Composers, Authors and Publishers ("ASCAP") and Broadcast

Music, Inc. ("BMI") consent decrees.[1] A true and correct copy of my curriculum vitae is attached as Exhibit A.

B. Assignment and Materials Used

3. Counsel for Anthropic PBC has asked me to evaluate the effects of using Plaintiffs' song lyrics as training data for Anthropic's large language model ("LLM") on the potential market to license text and/or song lyrics to train LLMs. My analysis focuses exclusively on the use of song lyrics as inputs for training LLMs to generate new text, which I understand to be the primary and prevalent use of LLMs like Anthropic's signature product, Claude.

4. To perform my analysis, I have used materials provided in discovery[2], academic articles, and other public information.

II. SUMMARY OF OPINIONS

5. The claim that there is an incipient market for training licenses implicitly assumes that such a market is economically viable. Economic analysis shows that the hypothetical competitive market for licenses covering data to train cutting-edge LLMs would be impracticable. This conclusion is based on the following subordinate conclusions.

a) In a competitive market for licenses to train LLMs, the highest price that an artificial intelligence ("AI") company would pay is the increase in value of the model created by the addition of the licensor's content to the training dataset.

---

[1] ASCAP, BMI, SESAC, and GMR are performing rights organizations ("PROs") that aggregate and license public performance rights in music compositions controlled by publishers like the plaintiffs in this proceeding.

[2] I understand that Plaintiffs' have represented they will be providing "rolling" productions in connection with Anthropic's discovery requests related to the pending preliminary injunction motion. Moreover, on January 14, 2024, Plaintiffs produced about an additional 75 agreements and amendments that I have not had the opportunity to analyze in detail. I reserve the right to update my analysis and opinions in this declaration as necessary to reflect my ongoing review of Plaintiffs' production.

b) Incremental contributions to the value of the LLM by individual copyright owners would be small.

c) Copyright owners' opportunity costs of licensing for training purposes is zero or near zero.

d) The rates for licenses for data to train LLMs would be low in a hypothetical competitive market.

e) To successfully acquire data to train a cutting-edge LLM, AI firms would need to license with thousands of publishers and likely millions of copyright owners of material on the internet. The negotiation and administration costs of acquiring licenses in a hypothetical competitive market for licenses covering training data would be high relative to the value of licensors' individual contributions.

f) Each parties' uncertainty regarding the value of a license for training data to the other will contribute to the failure of some negotiations, raising negotiating costs.

6. Given the vast amount of information required to train LLMs and the practical difficulties associated with licensing enough individual works to amass such information in a competitive market, LLMs would likely not exist if AI firms were required to license the works in their training datasets.

7. The agreements that Plaintiffs' expert Professor Smith adduces as evidence of AI firms licensing content either do not cover data to train LLMs or appear to provide consideration other than training rights. Even when taken as presented by Professor Smith, the existence of these agreements does not indicate that there is an incipient market for licenses covering data to train LLMs.

III. ANTHROPIC AND THE CLAUDE MODEL

8. Anthropic is an AI company that develops AI models known as LLMs.[3] LLMs process large amounts of text in order to "understand" how words work together and use that "learning" to produce text of its own, typically in response to a user-generated prompt. When developing an LLM, engineers first create a "neural network," a type of software that analyzes large datasets and extracts statistical information from language to better understand how language works.[4] LLMs "learn" how to recognize, interpret, and generate language through a "training" process that enables them to identify relationships and patterns between words.[5] Anthropic's LLM, called "Claude," is trained on data compiled from the Internet, non-public datasets that are commercially obtained, data from users and service providers, and internally generated data.[6] Anthropic estimates that the current version of Claude, known as Claude 2, was trained on a dataset of billions of texts.[7] These texts contained roughly [redacted] tokens, which are fragments of words in the texts, and are used in the training process. The [redacted] tokens are roughly

---

3 *See,* https://www.anthropic.com/ (accessed 1/5/2024) ("Anthropic is an AI safety and research company based in San Francisco. Our interdisciplinary team has experience across ML, physics, policy, and product. Together, we generate research and create reliable, beneficial, AI systems.")

4 Declaration of Jared Kaplan in Support of Defendant's Opposition to Plaintiffs' Motion for Preliminary Injunction ("Kaplan Declaration"), ¶¶ 13-16.

5 Kaplan Declaration, ¶ 16; "Public Comments of Anthropic PBC," Before the United States Copyright Office, Notification of Inquiry Regarding Artificial Intelligence and Copyright, October 30, 2023 ("Anthropic Comments"), p 6.

6 Anthropic Comments, p. 5.

7 Kaplan Declaration, ¶ 22.

equivalent to roughly [redacted] words.[8] [redacted].[9]

9. Training datasets are large because the goal of training is for the model to identify patterns and relationships between words. Therefore, training depends on the volume and diversity of text.[10] Testing shows that LLMs trained on larger volumes of data were more adept at the generative tasks (such as creating new expressions).[11]

10. Once trained, LLMs are able to perform a range of natural language processing tasks.[12] According to Anthropic, "Claude is a next-generation AI assistant."[13] Claude's use cases include: customer service,[14] parsing legal documents,[15] general advice or coaching,[16] summarizing

---

8 See, *e.g.*, "What are tokens and how to count them?" OpenAI, available at https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (accessed 1/14/2024).

9 Kaplan Declaration, ¶ 43.

10 Kaplan Declaration, ¶¶ 18-19.

11 "Tracing Model Outputs to the Training Data," August 8, 2023 ("[P]atterns of generalization became more abstract with model scale.") available at https://www.anthropic.com/index/influence-functions (accessed 1/14/2024).

12 *See* Kaplan Declaration, ¶¶ 9, 19.

13 https://www.anthropic.com/product (accessed 1/5/2024). Anthropic emphasizes its Claude models' abilities to perform content generation tasks, including "complex reasoning," "thoughtful dialogue," "coding," "detailed content creation," "text analysis," "summarization," and "document comprehension."

14 https://www.anthropic.com/product (accessed 1/5/2024) ("Claude ensures speedy and friendly resolution to customer service requests, saving costs and increasing satisfaction. Claude can also be taught when to hand off tasks to a human CSR, enabling your team to focus on the most complex challenges.").

15 https://www.anthropic.com/product (accessed 1/5/2024) ("Claude is able to parse legal documents and answer questions about them. Lawyers can reduce costs and focus on higher level work.").

16 https://www.anthropic.com/product (accessed 1/5/2024) ("Claude can be an always-available active listening companion for personal growth as well as career development, providing a space to listen or give advice.").

results from Internet search,[17] back-office tasks,[18] and sales.[19,20] In addition, analysis of how Claude is used reveals that users rely on Claude to edit, rewrite, and brainstorm on original writing projects, such as scripts, novels, and poems.[21] Notably, each of the use cases that Anthropic describes is *generative*—that is, Claude is not designed to simply regurgitate text or facts, but to use its "knowledge" to help users more efficiently access, understand, and convey information.[22]

IV. THERE IS NOT AN INCIPIENT COMPETITIVE MARKET TO LICENSE TEXT TO TRAIN LLMS

11. Plaintiffs allege that Anthropic's use of their text for training purposes "will irreparably damage Publishers' competitive position and their ability to negotiate future licenses with AI developers."[23] This claim presupposes the existence of a competitive market for licenses covering text used to train LLMs. However, Plaintiffs have produced no evidence that there is an

---

17 https://www.anthropic.com/product (accessed 1/5/2024) ("Claude is able to integrate seamlessly into web search as well as private search over knowledge bases, synthesizing search results triggered by user questions into natural language answers.").

18 https://www.anthropic.com/product (accessed 1/5/2024) ("Claude is able to handle a wide range of rote office work. It can extract relevant information from emails and documents, categorize and summarize survey responses, and generally wrangle reams of text with high speed and accuracy.").

19 https://www.anthropic.com/product (accessed 1/5/2024) ("Claude can act as an always-on and enthusiastic virtual sales representative, answering customer questions and guiding them towards products that meet their needs. Customize Claude with your brand's personality and tone.").

20 *See also,* Anthropic Comments, p. 2 ("Claude tends to perform well at general, open-ended conversation; search, writing, editing, outlining, and summarizing text; coding; and providing helpful advice about a broad range of subjects…*Claude is designed to serve as a creative companion, to enable people to produce new works* (emphasis added)."); Anthropic Comments, pp. 3-4, describing examples of productivity-enhancing uses of Claude since its launch.

21 Kaplan Declaration, ¶ 10.

22 Kaplan Declaration, ¶ 28.

23 Memorandum of Law in Support of Plaintiffs' Motion for Preliminary Injunction, November 16, 2023, p. 26.

existing or incipient competitive market to license training materials. More importantly, they have not shown that such a market could exist.[24]

12. With no evidence of an existing market to license text to train LLMs, the claim that there is an incipient market for text to train LLMs is baseless without analysis of the characteristics of the *hypothetical competitive market* to license training material. The economics of the hypothetical competitive market to license training data indicate that the market would be impracticable and would not be a viable means for AI firms to acquire the vast (and growing) amount of text required to train cutting-edge LLMs.

13. For such a hypothetical competitive market to license training data to develop and function, the benefits to market participants must exceed the costs that they bear transacting in the market. More specifically, AI firms accessing training data through the hypothetical market would have to be economically viable when acquiring training content through the hypothetical market. This requirement implies, in particular, that AI firms could acquire the quantities of text required to train the current generation of LLMs and continue to increase LLMs' capabilities through the hypothetical market *without infringing*.

14. The economic analysis below shows that the hypothetical competitive market to license data for training would have a huge number of licenses between AI firms and copyright owners, such as book authors. To the extent individual posters on the Internet can assert copyright

---

[24] In particular, I and my staff have reviewed the licenses that Plaintiffs produced in this matter [redacted].

claims over their posts, the number of licenses needed would be in the millions.[25] The small contribution of each licensor's content to the value of the LLM would lead to low license rates for even large licensors. Nevertheless, the cost to negotiate and manage the required licenses would be high, even assuming all the needed copyright owners could be identified. In addition, licensees' and licensors' uncertainty regarding the contribution of the data being licensed to the value of the LLM would lead to some negotiations failing, raising negotiation costs and reducing the availability of training material. Such a market would not be a practicable means for AI firms to acquire text for training purposes.

15. The few agreements that Plaintiffs' expert Professor Smith describes are not relevant to the licensing market at issue and do not support the conclusion that there is an incipient market to license training data (text) for LLMs. And they do not demonstrate the viability of a market to license training data were such licenses required for all copyrighted works in LLM training datasets.

A. The Hypothetical Competitive Market to License Training Data

1. A Hypothetical Competitive Market for Licenses Covering Training Data

16. This section describes the characteristics of competitive input markets to establish an economic framework in which a competitive market to license training data would function.

17. Training data is an input into the development of LLMs. In competitive markets for inputs, the price—or in this case the license rate—for each input is equal to the incremental

---

[25] *See, e.g.*, "Are Tweets Protected by Copyright?," Copyright Alliance, available at https://copyrightalliance.org/faqs/tweets-protected-copyright/#:~:text=As%20for%20who%20owns%20the,on%20or%20through%20the%20Services. (accessed 1/13/2024).

value that it creates (*i.e.,* the marginal revenue product).[26] In the context of negotiations to license training data from copyright owners, this principle implies that the highest amount an AI firm would be willing to pay to add a copyright owner's data to its training database is the increase in the value of the model that results from the addition of that particular training data. Real-world negotiations would settle on license rates lower than the AI firm's willingness to pay for the data and higher than the lowest amount that the licensor would accept to grant the license. In a workable competitive market, each copyright owner must anticipate that it is licensing content to train an AI model that *will exist whether the individual copyright owner licenses or not.* In this circumstance, each licensor's cost of granting the license, including opportunity costs, is essentially zero because most uses of the model will be in areas unrelated to the licensed content and the contribution of the copyright owner's content to the model has no discernable effect on the operation of the model.[27]

18. Many existing training datasets come largely from sources on the Internet.[28] The copyrights in existing training datasets containing text from the Internet are not well documented, which makes licensing the data in existing datasets impracticable or effectively impossible.[29] This

---

[26] *See* "Principles of Microeconomics" Seventh Edition, N. Gregory Mankiw p. 377, ("Economists sometimes call this column of numbers the firm's *marginal revenue product*: It is the extra revenue the firm gets from hiring an additional unit of a factor of production.")

[27] Each individual copyright owner's contribution is small and content in the training corpus is fungible. Kaplan Declaration, ¶ 19.

[28] *Jesse Dodge, et al.*, "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," Paul G. Allen School of Computer Science & Engineering, University of Washington, September 2021 (arXiv:2104.08758v2).

[29] *Shayne Longpre, et al.*, "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI," Data Provenance Initiative, November 2023, (arXiv:2310.16787v3**),** p. 2; *Jesse Dodge, et al.*, "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," Paul G. Allen School of Computer

problem is potentially still more difficult to address when multiple owners share ownership of a copyright. For example, co-ownership is a well-documented challenge when it comes to licensing musical compositions. Owners of musical compositions and their agents have been vocal about their practice and expectation that licensees procure licenses from all co-owners of musical compositions, meaning that several licenses may be required to secure the required rights to one song.[30]

19. If scraping data from the Internet and matching it to licenses is unworkable, then training databases would be limited to affirmatively licensed data provided by licensors. Thus, the competitive market for licenses covering training data would entail each AI firm affirmatively licensing its training data from individual known copyright owners, which could be publishers that own catalogs of copyrighted works or individual copyright owners, such as book authors, bloggers, or creators of social media.

20. As described below, negotiating and administering the large number of licenses required to train a cutting-edge LLM would be costly. There would be no cost-effective way to reliably identify and contact all of the copyright owners to license their works, given the tiny contribution of each copyright owner's work to the value of the LLM. In fact, it may be impossible to identify and contact the copyright owners of many works because a work may be "orphaned," meaning its copyright owner cannot reliably be identified or contacted even with diligent effort.[31]

Science & Engineering, University of Washington, September 2021 (arXiv:2104.08758v2).

30 Comments by the National Music Publisher's Association Submitted in Response to U.S. Department of Justice Antitrust Division September 22, 2015, Solicitation of Public Comments Regarding PRO Licensing of Jointly Owned Works, available at https://www.justice.gov/atr/public/ascapbmi2015/ascapbmi22.pdf.

31 "Orphan Works and Mass Digitization," United States Copyright Office, June 2015, p. 1, available at https://www.copyright.gov/orphan/reports/orphan-works2015.pdf ("a user's

Taken together, these considerations imply that an AI firm could not successfully (*i.e.*, comprehensively) license an LLM's entire training dataset in a competitive market for licenses.

21. Transaction costs might be reduced if copyright owners banded together to negotiate licenses. However, if multiple licensors were to combine to negotiate as a group, that group would be a cartel. The impact of the cartel on prices would depend on its size relative to the size of the training data and the ability of AI firms to be successful if they did not have a license from the cartel. In any event, cartels are inconsistent with a hypothetical *competitive* market to license training data.

2. Training Requires Large Amounts of Data Relative to the Size of Even a Large Publisher's Archive; Each AI Firm Would Need Thousands of Licenses

22. The hypothetical competitive market to license training data would require thousands of negotiations by each AI firm with large and small publishers – and potentially millions of negotiations with individual owners of material on the Internet.[32]

23. LLMs such as Claude and ChatGPT require large amounts of text for training. The most recent version of Claude was trained on approximately [redacted].[33] [redacted]

---

ability to seek permission or to negotiate licensing terms is compromised by the fact that despite his or her diligent efforts, the user cannot identify or locate the copyright owner. Second, in the case of mass digitization – which involves making reproductions of many works, as well as possible efforts to make the works publicly accessible – obtaining permission is essentially impossible, not necessarily because of the lack of identifying information or the inability to contact the copyright owner, *but because of the sheer number of individual permissions required*. (emphasis added)"). *See also* p. 38 ("Studies of library collections of printed, published books and similar works estimate that between 17% and 25% of published works and as much as 70% of specialized collections are orphan works.").

[32] Anthropic Comments, pp 9-10.

[33] For reference, one trillion is equal to one million multiplied by one million, or 10 raised to the power of 12 (*i.e.*, 10^{12}).

[redacted].[34] Press reports indicate that OpenAI's most recent version of Chat GPT was trained on approximately 13 trillion "tokens," [redacted] tokens used in the training dataset used to train the current version of Claude.[35]

24. The economic consequence of the size of training datasets is that they are significantly larger than the archives of even the largest publishers.

a) **NYT Archive.** If the *New York Times* averages, say, 150,000 words each day, then a 100-year archive of the *Times* would contain just under 5.5 billion words – less than [redacted] of the words used to train the current version of Claude. Put differently, more than [redacted] archives of the size of the *Times* would be needed to compile a text dataset equal to the size of the Claude training dataset, and still more would be needed to train the next version of Claude.

b) **Song Publishers.** The larger song publishers among the Plaintiffs claim to have between one million and 5.5 million songs in their catalogs.[36] Other song

---

34 Kaplan Declaration, ¶ 43.

35 "GPT-4 architecture, datasets, costs and more leaked," The Decoder, July 11, 2023, available at https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/ (accessed 1/6/2024). A token is a group of letters that is typically shorter than the typical English word.

36 "The Three Major Music Publishers Now Own or Control Over 10 million Songs between Them (Kind of)" Music Business Worldwide, October 20, 2022, available at https://www.musicbusinessworldwide.com/the-three-major-music-publishers-now-own-or-control-over-10-million-songs-between-them-kind-of/#:~:text=Sony's%20global%20music%20publishing%20operation,more%20than%201%20million%E2%80%9D%20songs (accessed 1/7/2024). ("The 'kind of' disclaimer in our headline above is in there because these are the separate figures reported by each of the major music companies and their parent companies. What these numbers don't tell us is how many of these songs are *duplicates* i.e. [*sic*] are administered by two or even all three of the major publishers, due to them having multiple writers represented by different companies.").

publishers presumably have fewer songs in their libraries. If a typical song has 300 words,[37] then a large Plaintiff with a one-million song library could contribute 300 million words, or [redacted] of the database used to train the current version of Claude (assuming it owns the rights to enter such a license). More than [redacted] song publishers with one-million song libraries would be needed to create a database of the size used to train the current version of Claude. Licensing many songs could be still more administratively costly as the result of many songs' copyrights being fractionally owned by multiple writers, publishers, or investors.

c) **All the Books in the World.** Google estimated in 2010 that there were approximately 130 million books in the world.[38] These books may not all have fully unique text – that is, some may be different copies of the same play with different commentary, for example. However, if each book were unique and had on average 100,000 words, roughly the number of words in a typical modern novel, all the books in the world would contain 13 trillion words. [redacted]

[redacted]

---

37 *See*, *e.g.*, William Wier, "Words Words Words, Are excessive lyrics ruining pop music?" Slate, March 11, 2008, available at https://slate.com/culture/2008/03/are-excessive-lyrics-ruining-pop-music.html (accessed 1/14/2024).

38 *Leonid Taycher*, "Books of the world, stand up and be counted! All 129,864,880 of you," Google Book Search, August 5, 2010, available at https://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html (accessed 1/7/2024) (Discussing the difficulty of identifying unique books, "One definition of a book we find helpful inside Google when handling book metadata is a 'tome,' an idealized bound volume. A tome can have millions of copies (e.g., a particular edition of 'Angels and Demons' by Dan Brown) or can exist in just one or two copies (such as an obscure master's thesis languishing in a university library). This is a convenient definition to work with, but it has drawbacks. For example, we count hardcover and paperback books produced from the same text twice, but treat several pamphlets bound together by a library as a single book.").

[redacted][39].

25. These calculations show that, if a hypothetical competitive market to license data for training LLMs were to exist, Anthropic and each of the other companies building LLMs would need to license data from thousands of large publishers. When the fact that many copyrights are held by individual authors[40] and authors' individual copyright interests in blog posts and other material on the internet are considered, AI firms would need to license content from millions of individual copyright owners to acquire the data needed to train a cutting-edge general LLM, without infringing. Moreover, the copyright owners of a substantial number of works cannot be identified even with a diligent search. And the next generation of LLMs is expected to be trained on still larger datasets, increasing the complexity of the licensing problem.

3. License Rates for Training Data Would Be Low

26. Competitive rates for licenses covering training data would be low in a hypothetical competitive licensing market. Training LLMs depends primarily on the quantity and diversity of text available.[41] The texts from different copyright owners used to train LLMs, such as Claude, each individually make a tiny contribution to the whole and, in addition, are fungible.[42] Put

---

39 *See* Kaplan Declaration, ¶ 23.

40 Book publishers, for example, may not control the rights required to provide training licenses, and negotiations with individual copyright owners would be needed to acquire rights to their works. Lucio Lanucara, "Who Owns the Copyright to Published Works?" The Michelson Institute for Intellectual Property, September 27, 2021, available at https://michelsonip.com/who-owns-the-copyright/ (accessed 1/13/2024) ("In the United States, the Copyright Act (Title 17 US Code) states that intellectual property belongs to the author, unless otherwise specified in a publishing contract.").

41 Kaplan Declaration, ¶¶ 18-29.

42 Kaplan Declaration, ¶ 19.

differently, an AI firm could exclude one copyright owner's data from a training set, or exclude the copyright owner's data and replace it with alternative data, without materially affecting the value of the LLM.[43] This implies that AI firms place a low value on incremental data for their training datasets – that is, they have a low willingness to pay for training data at the margin.

27. As described above, in a competitive market to license training materials, the negotiated license rate between an AI firm and a licensor would be less than the anticipated increase in the value of the LLM resulting from the addition of the licensor's data to the training dataset. As a result, license rates in the hypothetical competitive market would be low.

28. As described above, a Plaintiff with a 1 million song catalog could offer Anthropic [redacted] of the content needed to train the current version of Claude [redacted]. The incremental value of the content held by a large music publisher is small compared to the quantities of data that could be obtained from other publishers, but even they have too little data to justify opening negotiations. For example, OpenAI reports that it explained to *The New York Times* in the course of their negotiations "that, like any single source, their content didn't meaningfully contribute to the training of our existing models and also wouldn't be sufficiently impactful for future training."[44] Individual authors with copyrights in books could license still smaller shares of the data required to train an LLM. Moreover, the interchangeability of copyright owners' data puts the copyright owners in competition with each other to obtain licenses. With copyright owners facing competition among

---

[43] Kaplan Declaration, ¶ 19; *see* OpenAI and journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 1/13/2024).

[44] "OpenAI and journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 1/13/2024). OpenAI describes its negotiations with the times as not addressing training uses. Negotiations were "focused on a high-value partnership around real-time display with attribution in ChatGPT[.]"

themselves and facing opportunity costs that are essentially zero for training uses of their content, negotiated license rates would be low.

4. Negotiation and Transactions Costs in the Hypothetical Market Would Be High and Would Be Elevated by Uncertainty Regarding the Value of a Licenses

29. Negotiating licenses to acquire training data would be administratively costly. The costs involved entail at least the cost of reaching agreement and the costs of continuing to track content and administer licenses.

30. The *New York Times* negotiated with OpenAI for months regarding non-training uses of its data to train OpenAI's models.[45] Thus, both OpenAI (and Microsoft) and *The New York Times* apparently invested substantial time in their effort to reach an agreement – and still failed. If *The New York Times* does not have sufficient data to make it worthwhile for OpenAI to expend the cost and effort to negotiate a license to train using *The Times'* archives, the vast majority of publishers would not have content that is sufficiently valuable to justify the cost of entering into a training license. Moreover, any dataset large enough to be worth licensing has no value if the rest of the data needed to train the LLM is not available in the hypothetical competitive market.

31. Beyond the cost of negotiating licenses (*i.e.*, agreeing to the terms of the deal) is the administrative burden of tracking the content that licenses cover. Fox Corp. reportedly "launched a blockchain platform called Verify…to help media organizations monitor how their content is used online."[46] "Fox Corp. intends to use the Verify Protocol to negotiate deals licensing

---

45 Complaint, *The New York Times Company v. Microsoft Corporation, et al.,* December 27, 2023, ¶ 7.

46 Natalie Korach, "Fox corp. Launches Verify Tool to Check Authenticity of Content, Negotiate with AI Firms," The Wrap, January 9, 2024, available at https://www.thewrap.com/fox-corp-verify-tool-ai-negotiate-media-companies/.

content from Fox networks to AI firms."[47] The implication is that Fox invested in a software tool in order to track what it has licensed to AI firms. The need for and benefits of a software tool indicate the complexity and cost of administering licenses. The costs of tracking licenses are both the cost of developing and maintaining the software tool and the cost of inputting data and managing the data in the tool (*e.g.*, updating records to reflect changes of ownership of content). Addressing the administrability of managing copyrights, Anthropic's Chief Science Officer also notes that excluding an ever-expanding list of song titles from Claude's training database would "be impossible to achieve" in the course of Anthropic's business. [48]

32. A further complication in the hypothetical competitive market to license training data is that the incremental contribution to the value of an LLM created by a particular publisher's data is known or verifiable for the parties to a hypothetical licensing negotiation. Uncertainty of this kind (*i.e.*, when negotiators do not know the value of a deal to their negotiating partner), makes negotiations more likely to fail even when there are gains from reaching an agreement.[49] Thus, poor information regarding the competitive value of a publisher's training data to an AI firm implies that negotiations could fail to reach agreement. Time spent on negotiations that do not lead to agreement would make the hypothetical negotiations for licenses more costly and reduce the amount of training data available through the competitive market.

---

47 Natalie Korach, "Fox corp. Launches Verify Tool to Check Authenticity of Content, Negotiate with AI Firms," The Wrap, January 9, 2024, available at https://www.thewrap.com/fox-corp-verify-tool-ai-negotiate-media-companies/.

48 Kaplan Declaration, ¶ 45.

49 *See* "Microeconomic Theory", Andreu Mas-Colell, Michael Whinston, and Jerry Green p. 895 ("Whenever gains from trade are possible, but not certain, there is *no* ex post efficient social choice function that is both Bayesian incentive compatible and satisfies these interim participation constraints."

33. In short, given the vast amount of information required to train LLMs and the practical difficulties associated with licensing enough individual works to amass such information, LLMs would likely not exist if required to license the works that make up their training datasets. The need to negotiate with large numbers of copyright owners, many with only small contributions to a training database, implies that the costs of licensing would exceed the incremental benefits conveyed by the licenses in a hypothetical competitive licensing market for LLM training materials.

B. The Agreements Plaintiffs Have Identified Do Not Indicate that There Is an Incipient Market to License Training Data

34. The hypothetical competitive market for training data would entail thousands of licenses for each AI firm. The existence of a few licenses—which address non-text data for image or music AI models and may be driven by idiosyncrasies of licensors and differences in the amounts of training data required for those products—does not demonstrate the practicability of a competitive market to license data to train LLMs.

35. Plaintiffs' expert, Professor Smith, states that "several AI companies have publicly entered into license agreements concerning a variety of other forms of creative content."[50] Notably, Professor Smith does not claim to have identified licenses covering text for the purpose of training LLMs. Nor does he appear to have reviewed the agreements themselves. The licenses that Professor Smith enumerates either are not related to training LLMs or their terms are not clear. As a result, they do not support an inference that there is an incipient market to license all copyrighted textual data required for training LLMs.

---

[50] Declaration of Michael D. Smith in Support of Plaintiffs' Motion for a Preliminary Injunction ("Smith Declaration"), 11/17/23, ¶ 28.

a) **Stability AI.** Stability AI trained its audio generation model with data from AudioSparx and claims that by partnering with a licensing company it "has permission to use copyrighted material."[51] The article that Professor Smith cites does not describe the nature of the partnership or the license, and specifically does not address whether it covered training. Moreover, the audio generator appears to have been trained on a single dataset of audio recordings, which indicates the data requirements required to train AudioSparx are far lower than the requirements faced by AI firms building cutting-edge LLMs. Thus, this example is not reflective of the burdens of licensing sufficient data to train an LLM.

b) **Generative AI by Getty Images.** Getty worked with Nvidia to launch Generative AI by Getty Images. "Generative AI by Getty Images (yes, it's an unwieldy name) is trained only on the vast Getty Images library, including premium content, giving users full copyright indemnification."[52] This example does not provide information relevant to licensing text from third parties to train LLMs. First, Getty trained the image generator *using only its own data*. Thus, the volume of information needed to train the AI model to generate images appears to be much lower than in the case of LLMs. Second, Getty assures users that they have copyright protection, which indicates the possibility that *output* from the model could infringe one of Getty's images. Any licensing or other

---

51 Emilia David, "Stability AI Releases AI Audio Platform", The Verge, September 13, 2023, available at https://www.theverge.com/2023/9/13/23871635/stability-ai-generative-audio-model-platform (accessed 1/7/2024).

52 Emilia David, "Getty made an AI generator that only trained on its licensed images, The Verge, September 25, 2023, available at https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch (accessed 1/7/2024).

considerations regarding potentially infringing output are separate from the issues raised using data for training.

c) **OpenAI-Associated Press.** OpenAI has licensed news stories from Associated Press. According to OpenAI and Associated Press "'The arrangement sees OpenAI licensing part of AP's text archive, while AP will leverage OpenAI's technology and product expertise[.]'"[53] The description of the license does not indicate that OpenAI is paying to train its LLMs using the Associated Press's articles. In fact, the description of the deal is substantially broader, with direct benefits to AP's efforts to use AI.

d) **OpenAI-Shutterstock.** OpenAI has licensed images[54] and agreed to a license for "access to additional Shutterstock training data including Shutterstock's image, video and music libraries and associated metadata."[55] Thus, the agreement to access Shutterstock's data covers images, videos, music, and their associated metadata. The license is not reported to cover a text database. As described above, even if the licenses explicitly covered use of the images and music for training, these licenses are not informative about the hypothetical market for text data to train LLMs.

---

53 Matt O'Brien, "ChatGPT-maker OpenAI signs deal with AP to license news stories," July 13, 2023, available at https://perma.cc/HZF6-S7KF (accessed 1/7/2024).

54 "Shutterstock Partners with OpenAI and Leads the Way to Bring AI-Generated Content to All," October 25, 2022, available at https://www.shutterstock.com/press/20435 (accessed 1/7/2024).

55 "Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data," July 11, 2023, available at https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year (accessed 1/7/2024).

36. As suggested by the examples above, there are several reasons that AI firms would "license" data. **First**, AI firms may license data for purposes that are not inconsistent with their beliefs that training LLMs with copyrighted data is fair use. Indeed, OpenAI has publicly maintained its position that training LLMs with copyrighted data is fair use.[56] As described in the article discussing the license between OpenAI and Shutterstock, OpenAI licensed access to "Shutterstock data." However, the article does not state that training uses are licensed. I understand that AI firms have purchased *access* to data to make it available for training without specifically licensing the underlying copyrights for training.

37. **Second**, some licenses with copyright owners supplying data to AI firms allow for the output of the AI firm's models to reproduce the training data to a degree that is beyond fair use. For example, OpenAI recently "agreed to pay German media conglomerate Axel Springer, which publishes *Business Insider* and *Politico*, to show parts of articles in ChatGPT responses."[57]

---

[56] See, *e.g.*, *Silverman v. OpenAI, Inc.*, ECF No. 32, No. 23-cv-03416-AMO (N.D. Cal. Dec. 8, 2023) at 2-3; *Tremblay v. OpenAI, Inc.*, ECF No. 33, No. 23-cv-03223-AMO (N.D. Cal. Dec. 8, 2023) at 2-3; *see also* "OpenAI and Journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 1/13/2024) ("Training is fair use, but we provide an opt-out because it's the right thing to do"); OpenAI Letter Re: Notice of Inquiry and Request for Comment (Docket No. 2025-06), October 30, 2023, p. 11 ("OpenAI believes that the training of AI models qualifies as a fair use, falling squarely in line with established precedents recognizing that the use of copyrighted materials by technology innovators in transformative ways is entirely consistent with copyright law.")

[57] Will Oremus and Elahe Izadi, "AI's future could hinge on one thorny legal question," *The Washington Post*, January 4, 2024, available at https://www.washingtonpost.com/technology/2024/01/04/nyt-ai-copyright-lawsuit-fair-use/ (accessed 1/7/2024).

Thus, some licenses between AI companies and copyright owners cover uses other than training on publicly available content.[58]

38. **Third**, a license for access to data with a copyright owner that threatens to sue may be less expensive than defending a lawsuit, even if the AI firm expects to ultimately prevail.

39. In short, none of the licenses Professor Smith describes supports the conclusion that a hypothetical competitive market to license training data would be able to provide AI firms with sufficient data to train today's cutting-edge LLMs, such as Claude.

C. Using Song Lyrics to Train LLMs Does Not Threaten Plaintiffs' Existing Licenses

40. Plaintiffs' opportunity cost of licensing data for training uses is zero or nearly zero. Training uses input data to teach an LLM how to process language and create new outputs. Moreover, such uses do not threaten Plaintiffs' existing licenses. For example, licenses to existing lyric websites might be threatened by services that regurgitate a wide array of song lyrics over an extended period of time.[59] However, training Claude or another LLM using text data does not use lyrics, or any other works, for the purpose of regurgitating training data as outputs.[60] Claude's use of lyrics combined with other inputs to create new content does not threaten existing licenses with licensed lyric websites.

---

58 *See, e.g.*, "OpenAI and journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 1/13/2024) (explaining OpenAI's approach to licensing).

59 It would be cheaper and easier to obtain accurate song lyrics from a licensed site than to attempt to get an LLM to regurgitate a potentially incorrect or incomplete set of lyrics, assuming that the song of interest was "memorized" by the LLM and that potentially infringing outputs are not effectively filtered before presentation to the user.

60 Kaplan Declaration, ¶ 28.

I declare under penalty of perjury that to the best of my knowledge, information, and belief, the foregoing statements are true and correct.

Executed on January 16, 2024 at Arlington, MA.

Steven R. Peterson